SCALES: Scalable Analysis and Logging of Event Systems

Ke-Thia Yao, Bob MacGregor, Bob Neches
Bob Lucas, Dan Davis

University of Southern California
Information Sciences Institute

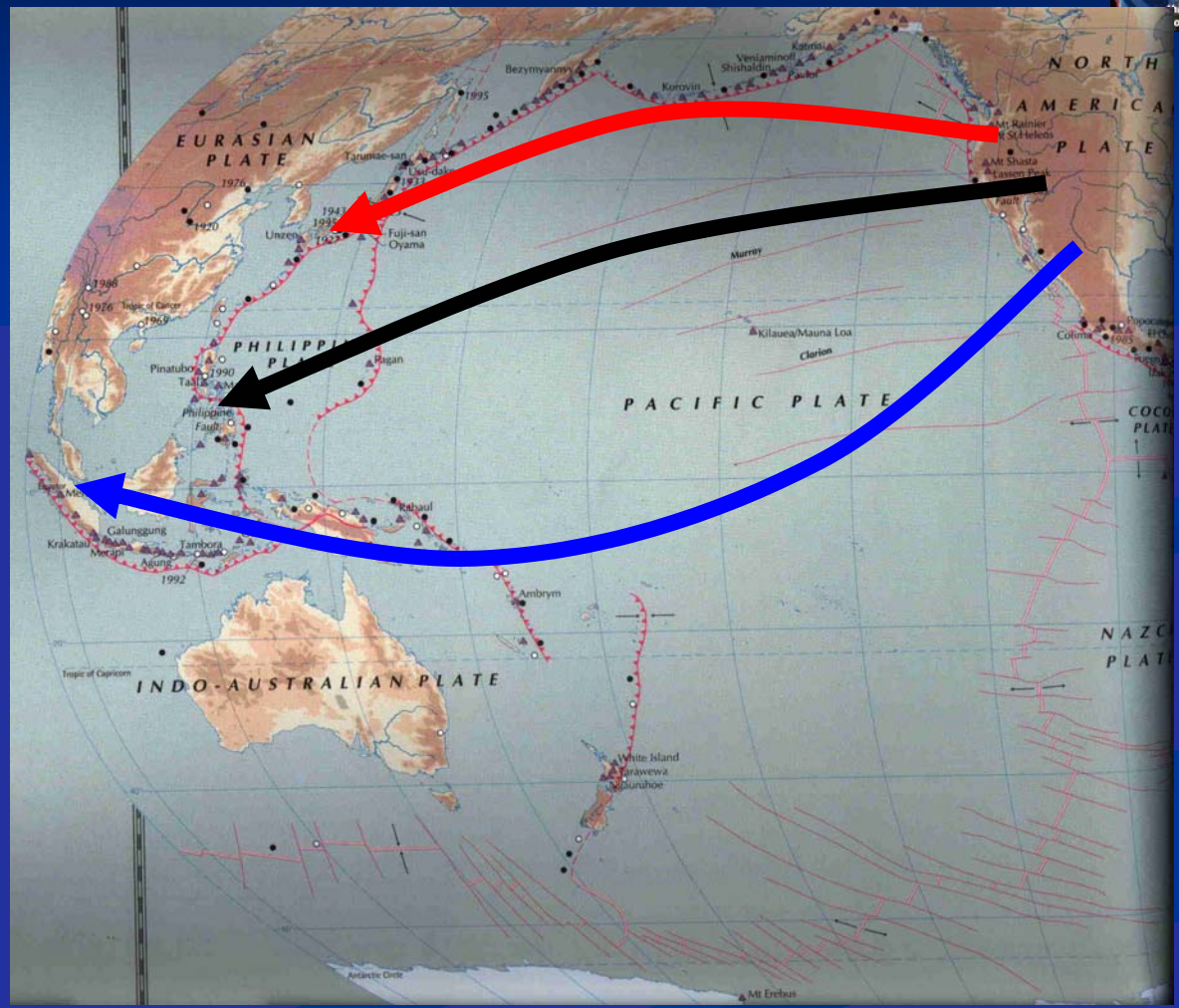
JFCOM J-9 Distributed Continuous Experimentation Environment

Playbox

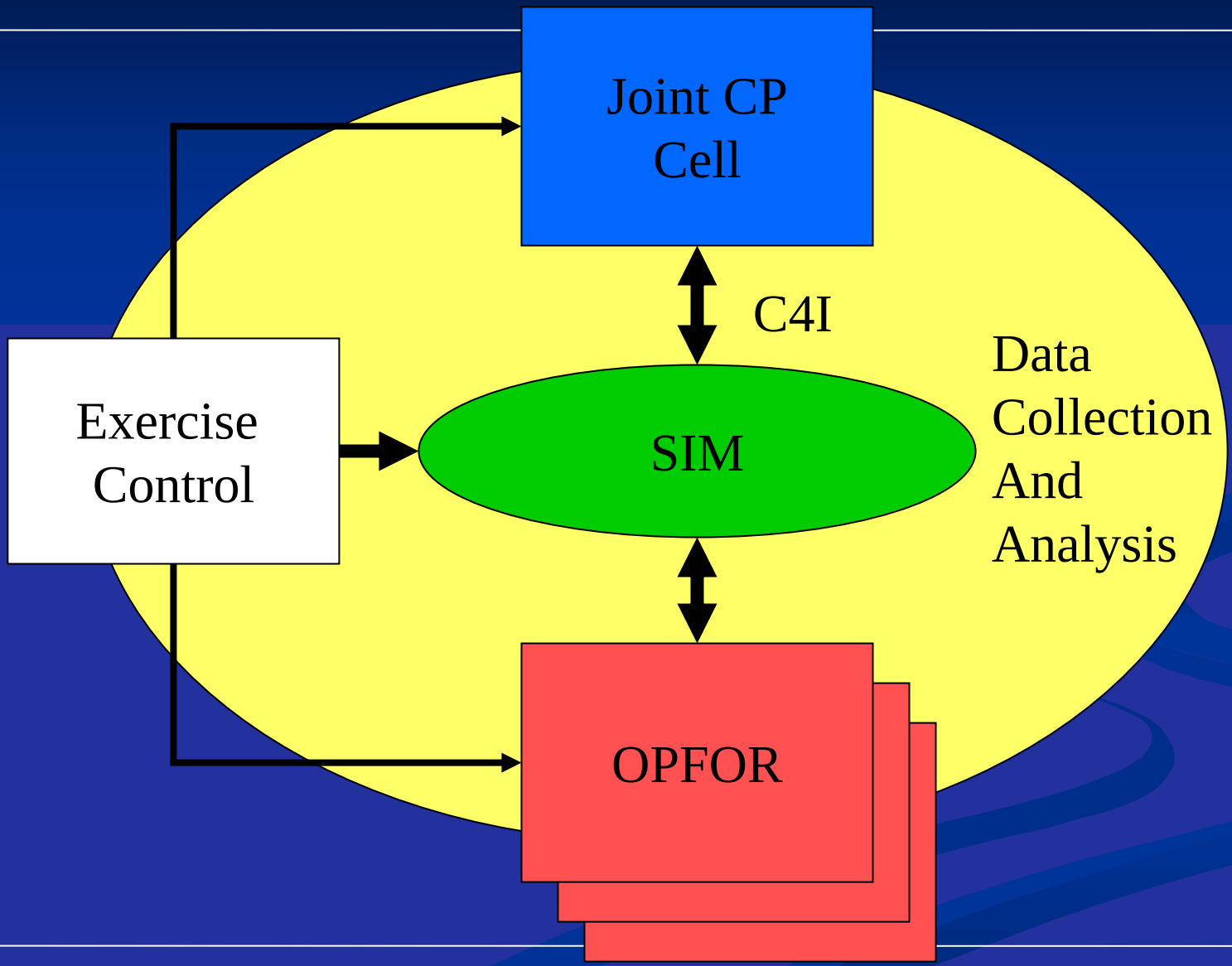
- Pacific Rim
- 105W-90E
- 50N-50S

Sample Scenario

- Terrorists take over parts of Philippines
- Disaster in Japan
- Dirty Bomb in Singapore



Basic Problem



Problems Exacerbated by Scale: 1+ Million Entities

- Sheer number of entities
  - Overwhelm communications and hardware
  - Overwhelm human operator's ability to simultaneously control entities
- Heterogeneous computing resources
  - Distributed
  - Multiple SPP with varying CPU, Memory, Bandwidth complicates resource utilization

Areas of Focus

- Simulation setup
  - Are there sufficient computing resources to support the desired simulation scenario?
  - How to define the initial condition of the simulation?
- Simulation visualization and understanding
  - How to monitor and analyze what the simulation is doing?
- Simulation control
  - How to adjust the simulation to keep it within expected bounds?

Maximally Effective Use of Simulation Data Logs for Analysis

SCALES: Data Logging and Analysis

Military Users

- Measure effectiveness: situation awareness, precision engagement/collateral damage, etc.
- Compare and contrast: e.g., evaluate simulation ground truth against sensor observations

➢ Near real time control; Quantify lessons learned

Simulator Developers

- Better debugging environment: e.g. check pointing and simulation restarts
- Check simulation events/patterns against expected behavior to find anomalous behavior

➢ Higher fidelity simulations

Infrastructure Managers

- Monitor CPU / memory / network resource usage, correlate with activity
- Discover faults and resource usage bottlenecks

➢ Higher fidelity battlefield monitoring; Larger, faster simulations

What We Are up Against: Million-entity Data Profile

- High data rate
  - *Selective* logging for analysis (FCS exercise)
    - 100 MB/hour for 20,000 entities (~500 non-clutter)
    - 1 million entities => 5 GB/hour
  - *Full* logging for playback
    - 2 GB/hour for 10,000 entities
    - 1 million entities => 200 GB/hour
- Huge amount of data
  - 1 million entities for 5 day event
    - *8 Terabytes*
    - ~5 days to transfer data using dedicated OC-3 line (155Mb/s)
    - ~2 weeks to transfer data using dedicated T3 line (45Mb/s)

Two Key Challenges

- Collect the “fire hoses” of data generated by large-scale distributed sensor rich environments
  - Without interfering with battlefield communication
  - Without interfering with simulator performance
- Maximally exploit the collected data efficiently
  - Without overwhelming users
    and without losing critical content

Target: unified distributed logging/analysis infrastructure,
helps military users and computing/networking infrastructure managers

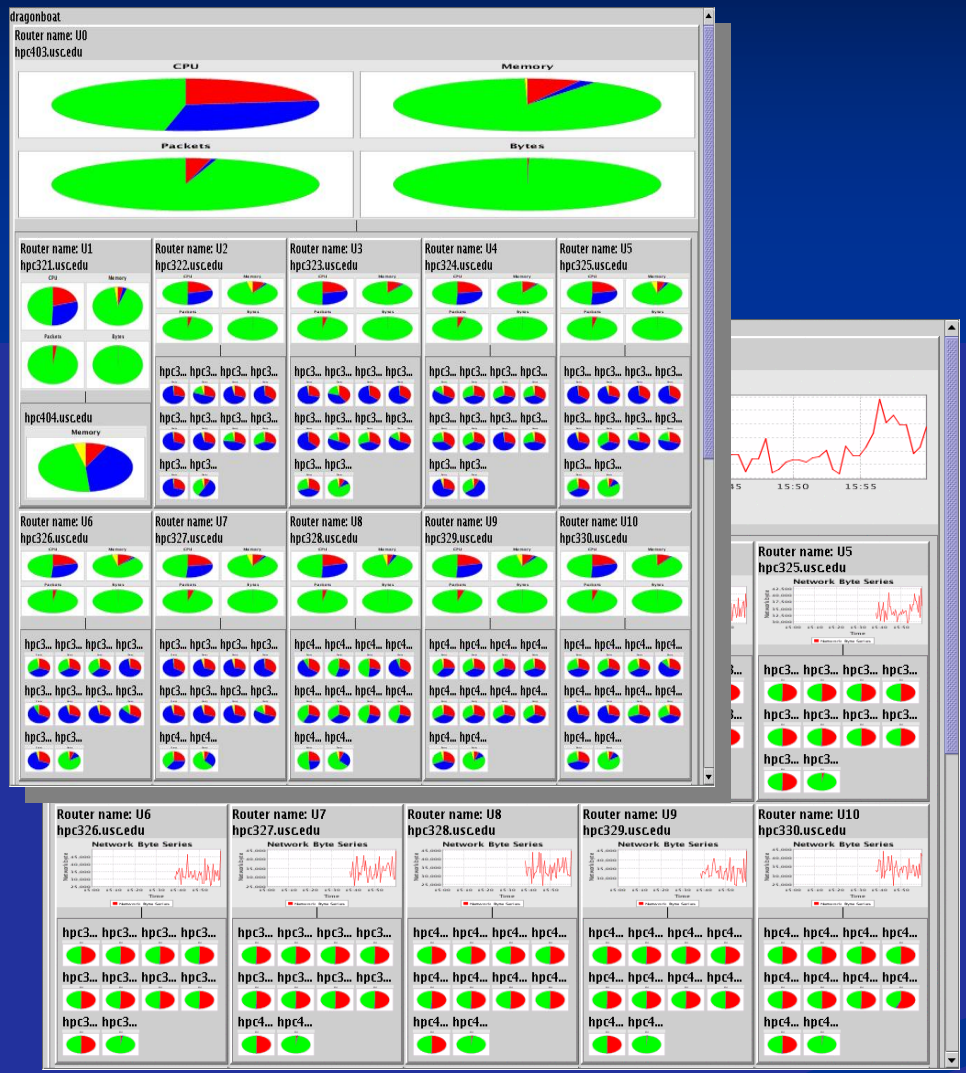
Approach

1. Provide better component metadata
   - Help designers express what they have created
   - Help other designers understand what they're working with
2. Provide metadata-level scripting mechanism
   - Help designers assemble software applications
3. Provide software gauges
   - Help application developers make component selections
   - Help component developers insert new components into the component database
   - Help system architects/administrators make application adaptations

Scalable, Minimally-Intrusive Real-Time Data Capture

- *SCALES solution: use parallelism at every phase*
  - *Minimize network communication overhead by*
    - Logging distributed data near point of generation
    - Selectively propagating data based on need
  - *Maximize use of computation resources by distributing analyses across light-weight DBMS's at each site*
  - *Multi-modal exploration engineered to work with distributed data to aid mining, analysis, and visualization*

MRI: Monitoring Remote Imaging



- Monitors resource usage of remote computing nodes
- Displays resource usage in context of connection architecture
- Stores data for post-mortem analysis

- **Types of resources monitored: CPU, memory, network traffic (bytes, packets)**
- **Display types: pie charts, time series**

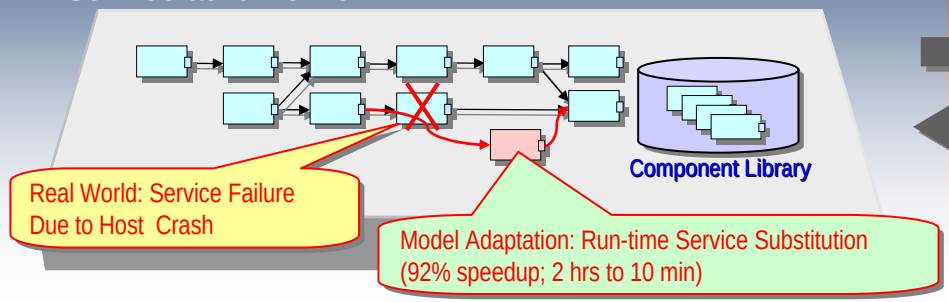
Sample Application: Monitoring distributed execution of over 1M entities distributed over 256 processors in JFCOM Joint Experimentation environment utilizing Simulated Autonomous Forces

Semantic Interoperability Measures: Using Multi-level Architecture Views to Overcome Faults/Bottlenecks

- Software architecture views enable dynamic, rapid response to faults by
  - Providing visibility into software systems
  - Identifying control points to adjust their behavior
- Multi-level views offer a greater range of adjustments than any single level
  - A system architecture view enables dynamic adjustment of servers:
    - Create additional server to accommodate increased demand
    - Migrate server from overloaded host to new host
  - A dataflow architecture view enables reformulation of an application:
    - Substitute alternative type of service for non-functioning or unavailable service
- Simulation and info. mgt. applications provide testbed for monitoring/repairing faults
  - GeoTopics “Hot News” Portal application executes as 120 individual components

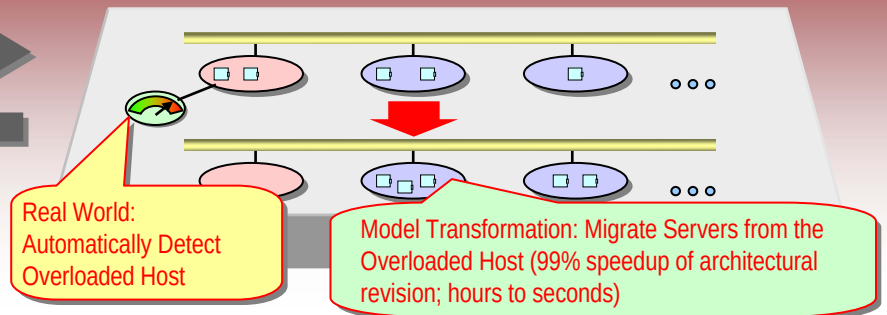
Dataflow Architecture View

- **Load dataflow architecture; extend it at run-time**
- **Update dataflow architecture to replace malfunctioning service at run-time**



System Architectural View

- **Detect overloaded server; re-host the service**
- **Update system architecture automatically to reflect re-hosted service**



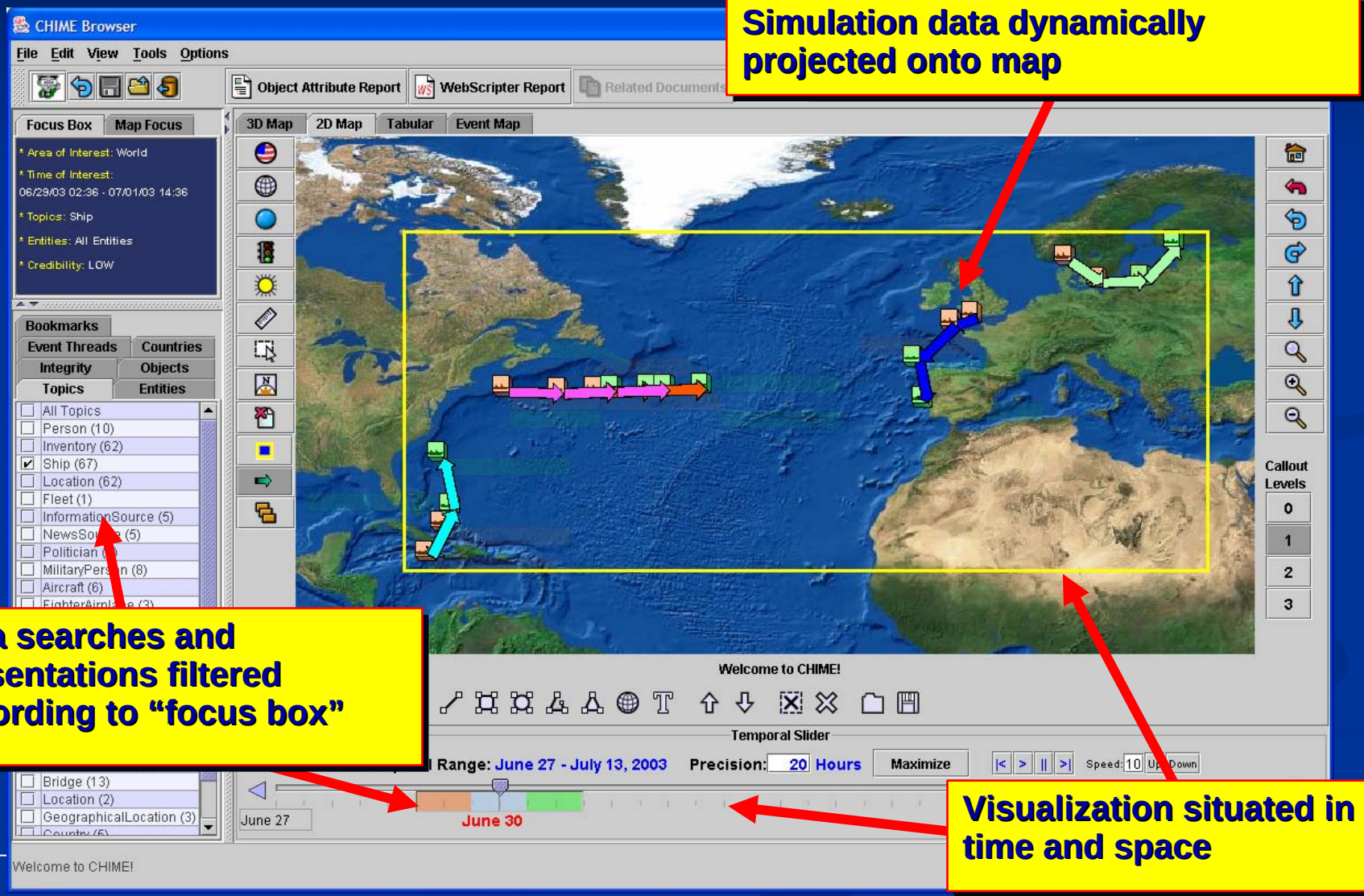
Multi-Modal Exploitation

- Coordinate alternative graphical presentations helps users understand data
  - Maps, tables, charts, time-based animations
- "Temporal peripheral vision" helps users notice potentially interesting events
  - Temporal "focus box" accentuates near-by events
- "n-Dimensional filtering" helps users home in on relevant data
  - E.g., narrowing temporal focus box and adding entity type constraints

N-dimensional Modeling Techniques Enable Mining, Analysis and Visualization

- Dimensions of interest intuitively slice data
  - E.g., geographical AOI, time, entity type, domain, echelon
- Conformed dimensions allow comparison across data sources
  - E.g., comparing simulation ground truth vs sensor observations with respect to same geographical AOI
- Aggregation aids cognitive grasp of larger data sets, improves query performance
  - E.g., summarize detonations at multiple grain sizes (country, state, county, day-of-week, hour, minute)

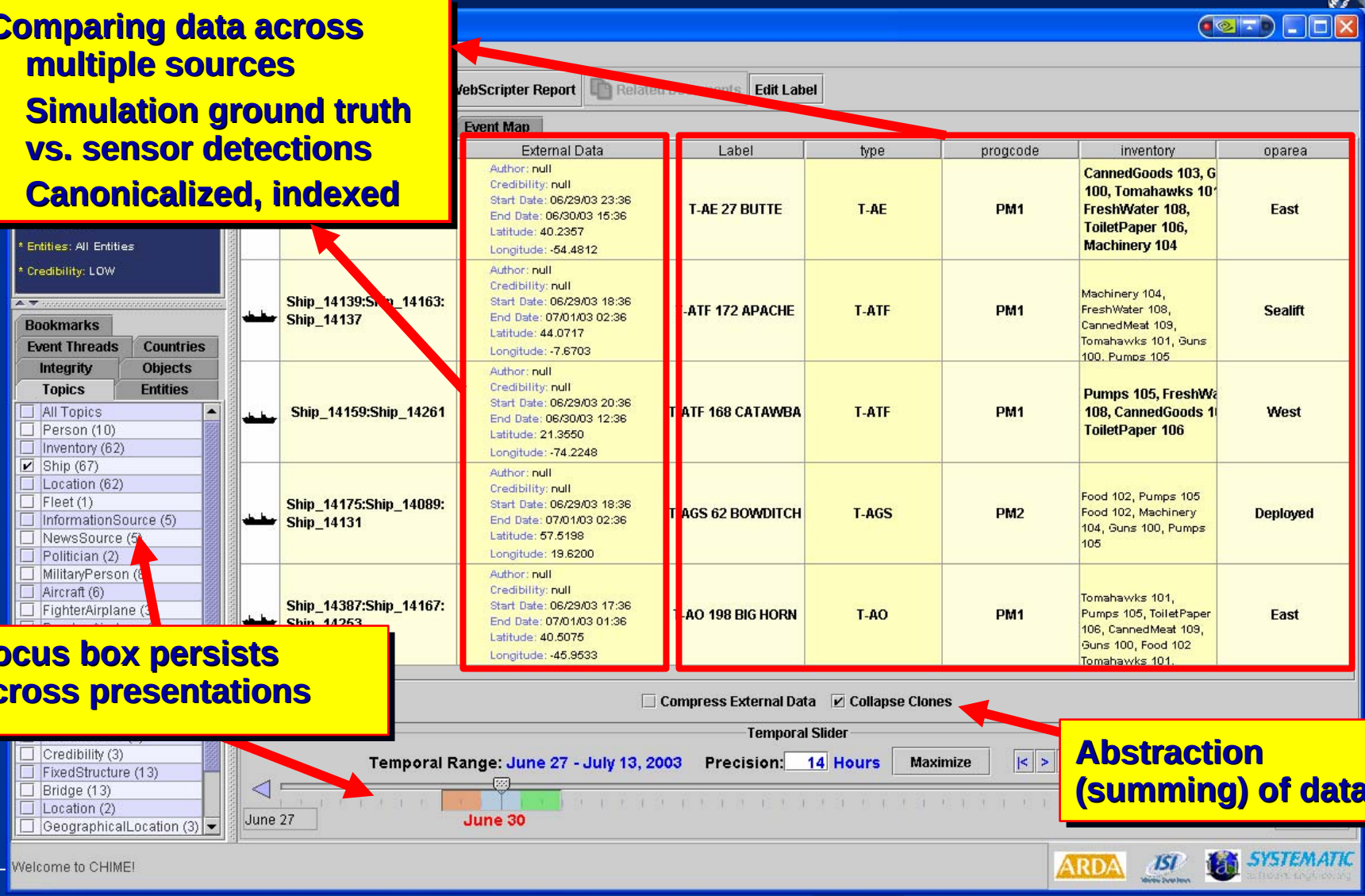
Multi-Modal Displays: Map-Based



Multi-Modal Displays: Tabular

Comparing data across multiple sources

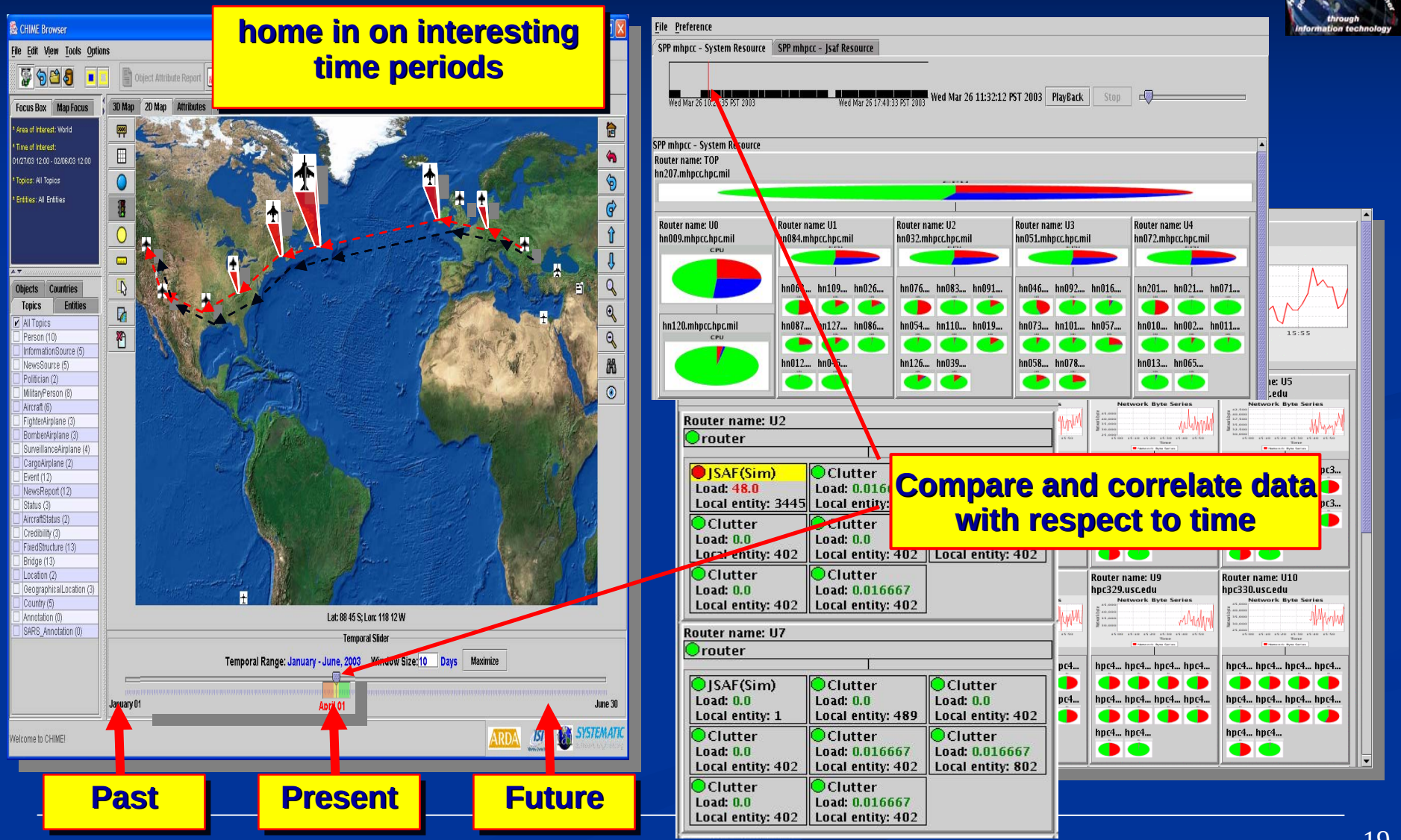
- **Simulation ground truth vs. sensor detections**
- **Canonicalized, indexed**



Focus box persists across presentations

Abstraction (summing) of data

Play and Review Simulation Performance



Benefits

- Operational Users:
  - Improved situational awareness and near real time control
  - Detailed after-action report / lessons learned (same whether derived from live action or training simulations)
- Acquisition Program Managers:
  - Use after-action / lessons learned to evaluate alternatives
  - Capture data from instrumented live action for evaluating impact of alternative technologies in future experiments
  - Test C4ISR systems before deployment
- Systems software support
  - Helps both live action and simulation infrastructure managers discover faults and computational bottlenecks
  - Helps simulation developers create *realistic simulations* by providing better debugging environments